land

MDPI

Article

Research on the Optimization of Multi-Class Land Cover Classification Using Deep Learning with Multispectral Images

Yichuan Li [1], Junchuan Yu [1,*], Ming Wang [1], Minying Xie [1], Laidian Xi [2], Yunxuan Pang [2] and Changhong Hou [3]

[1] China Aero Geophysical Survey and Remote Sensing Center for Natural Resources, Beijing 100083, China; lyichuan@mail.cgs.gov.cn (Y.L.); wangming@mail.cgs.gov.cn (M.W.); xieminying@mail.cgs.gov.cn (M.X.)
[2] School of Earth Sciences and Resources, China University of Geosciences (Beijing), Beijing 100083, China; 2101210160@email.cugb.edu.cn (L.X.); 2101210161@email.cugb.edu.cn (Y.P.)
[3] School of Geosciences and Surveying Engineering, China University of Mining and Technology (Beijing), Beijing 100083, China; bqt2300203041@student.cumtb.edu.cn
* Correspondence: yujunchuan@mail.cgs.gov.cn; Tel.: +86-15101157141

Abstract: With the advancement of artificial intelligence, deep learning has become instrumental in land cover classification. While there has been a notable emphasis on refining model structures to improve classification accuracy, it is imperative to also emphasize the pivotal role of data-driven optimization techniques. This paper presents an in-depth investigation into optimizing multi-class land cover classification using high-resolution multispectral images from Worldview3. We explore various optimization strategies, including refined sampling strategies, data band combinations, loss functions, and model enhancements. Our optimizations led to a substantial increase in the Mean Intersection over Union (mIoU) classification accuracy, improving from a baseline of 0.520 to a final accuracy of 0.709, which represents a 35.2% enhancement. Specifically, by optimizing the classic semantic segmentation network in four key aspects, we improved the mIoU by 15.5%. Further improvements through changes in data combinations, sampling methods, and loss functions led to an overall 17.2% increase in mIoU. The proposed model optimization methods enabled the OUNet to outperform the baseline model by providing more precise edge detection and feature representation, while reducing the model parameters scale. Experimental evidence shows that in the application of multi-class land surface classification, increasing the quantity and diversity of samples, avoiding data imbalance issues, is equally valuable for improving overall classification accuracy as it is for enhancing model performance.

Keywords: deep learning; multispectral; Worldview3; classification; model optimization

Citation: Li, Y.; Yu, J.; Wang, M.; Xie, M.; Xi, L.; Pang, Y.; Hou, C. Research on the Optimization of Multi-Class Land Cover Classification Using Deep Learning with Multispectral Images. *Land* **2024**, *13*, 603. https://doi.org/10.3390/land13050603

Academic Editor: Le Yu

Received: 24 March 2024
Revised: 27 April 2024
Accepted: 28 April 2024
Published: 30 April 2024

1. Introduction

High-resolution remote sensing plays a critical role in capturing detailed surface information and is widely applied in land use monitoring, environmental assessment, agricultural management, urban planning, and disaster response, providing precise data support for decision-making [1,2]. Over the past two decades, land cover classification based on remote sensing images has evolved through various stages. Traditional methods primarily relied on manual interpretation, which, despite its widespread use and high accuracy, was labor-intensive and limited by prior knowledge [3]. With the advent of multispectral and hyperspectral data, researchers have employed spectral indices to extract specific feature information for land cover classification, complemented by manual interpretation. For example, due to the high reflectance of vegetation in the short-wave infrared (SWIR) band, the spectral characteristics of SWIR are important for distinguishing features such as forests and water. However, these methods required manual feature extraction and threshold setting and often overlooked the spatial correlation of remote sensing features at the pixel level [4]. Subsequently, traditional machine learning methods such as Support Vector Machines (SVM) [5] and Random Forests [6], combined with object-oriented techniques,

became the mainstream for land cover classification, leveraging both spectral and spatial information from remote sensing data and addressing the issue of manual threshold setting to some extent [7]. However, the accuracy of these methods, particularly for linear features, was highly dependent on the initial multi-scale segmentation [8], leading to suboptimal performance in certain scenarios. For instance, researchers found that the standard object detection approach to segmenting remote sensing imagery did not accurately capture linear features such as rivers and roads.

The rapid development of artificial intelligence has led to significant advancements in machine learning methods for remote sensing applications [9,10]. Deep learning, in particular, has shown promise in land cover classification, offering higher accuracy and efficiency compared to traditional machine learning techniques and expert interpretation, especially when ample training samples are available [11,12]. Since the introduction of the first end-to-end semantic segmentation model, Fully Convolutional Networks (FCN) [13], in 2015, the field has seen rapid progress. Models like DeeplabV2 [14] have introduced atrous convolution to expand the receptive field of Convolutional Neural Networks (CNNs), while networks like PSPNet [15] have employed pyramid pooling to capture multi-scale features. The encoder-decoder architecture of UNet [16] and SegNet [17] has improved computational efficiency by reusing encoder features, and balancing shallow and deep features. DeepLabV3+ [18] further expanded this architecture, enhancing spatial resolution with powerful backbone networks and lightweight decoders. These models, along with others like DlinkNet [19], BiSeNet [20], HRNet [21], OCRNet [22] and Segformer [23], have propelled the field of semantic segmentation into a new era of development.

Leveraging the powerful learning capabilities of CNNs, it is possible to effectively capture contextual information and achieve recognition of land cover features within complex scenes [24]. Despite these successes, challenges remain in applying deep learning to multi-class target identification in remote sensing. Although remote sensing images are a type of natural image, the scale and morphology of the natural land features they depict often differ significantly from those of artificial features in high-resolution images [25]. Additionally, the background in remote sensing images is more complex than that found in typical natural images, with a larger proportion of information, which can lead to an imbalance between background and foreground details [26]. Furthermore, class imbalance in multi-class scenarios can lead to a reduction in overall classification accuracy [3]. For example, in wetland classification, the expansive water bodies often overshadow the smaller patches of marsh vegetation, leading to classification challenges. At present, in the research on multi-class land cover classification based on high-resolution remote sensing data, many researchers try to solve the problems of scale and morphological differences through multi-scale feature fusion and attention mechanisms [27–30]. However, there is relatively less emphasis on addressing the issue of imbalance problems. In addition, a trend can be observed where most studies focus more on improving classification accuracy by optimizing the model structure and less on mining the value of the data itself [31,32]. The purpose of this paper is to discuss the challenges of land cover classification using high-resolution remote sensing imagery, with a focus on the application of deep learning and addressing optimization issues like data imbalance and poor classification accuracy.

In this paper, experiments on remote sensing multi-class land cover classification were conducted based on Worldview3 data, and strategies to improve classification accuracy were proposed in terms of sampling methods, band combination, loss function, and model optimization. The purpose of our study is to verify that a reasonable optimization approach for training data is more important for remote sensing multi-classification than model optimization alone. All data and code used in this paper are open source in Github. Links are provided at the end of the document.

2. Data

The Worldview3 data used in this study was released during the Dstl Satellite Imagery Feature Detection competition [33], provided by the Defense Science and Technology

Laboratory of the United Kingdom. This is one of the earliest remote sensing classification datasets, which has laid an important foundation for the application of artificial intelligence in land cover classification [34]. The data have a spatial resolution of 1.24 meters, are preprocessed with geometric and radiometric correction, and consisting of 15 scenes of 1 km × 1 km multispectral data. Experts interpret satellite imagery to distinguish five land cover types: buildings, roads, trees, cultivated land, and water bodies. GIS software is utilized to digitize features, assign attributes, and generate a precise, labeled mask through visual interpretation as a reference to the actual ground conditions for accurate land cover classification. The remote sensing data was divided into two groups according to different band combinations: one group consisted of three bands for true color, while the other group consisted of seven bands, including yellow, red edge, and two near-infrared bands in addition to true color, covering a wider spectral range (Figure 1). The specific composition can be found in Table 1.

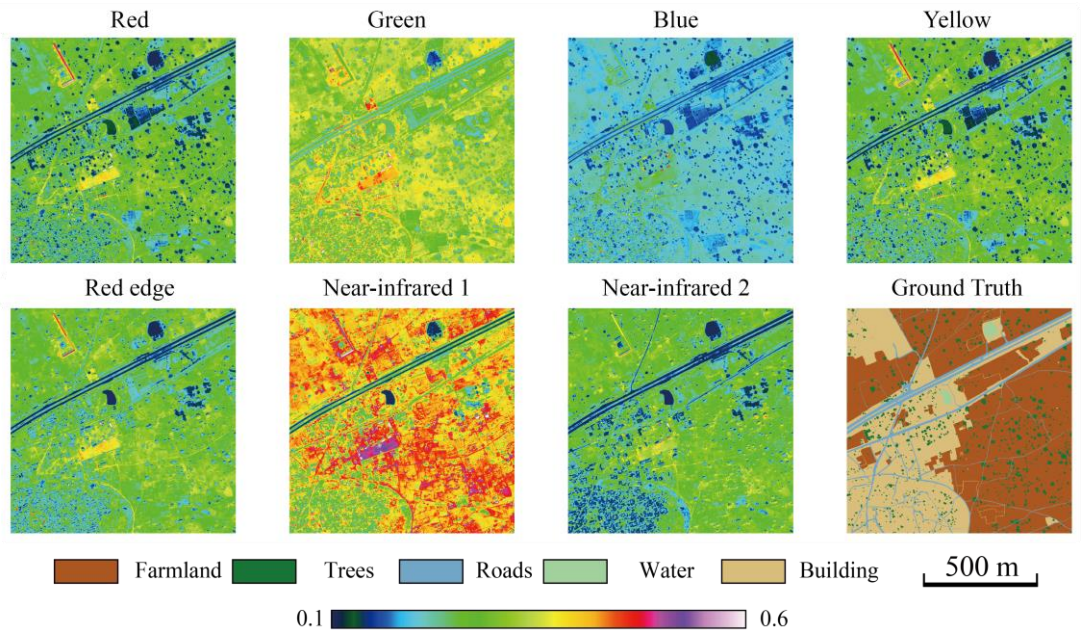

Figure 1. Composition of multispectral bands and classification labels.

Table 1. Band composition of worldview3 images.

Band Combinations	Band Number	Band Name	Wavelength (nm)
Three bands	1	Blue	450–510
	2	Green	510–580
	3	Red	630–690
Seven bands	1	Blue	450–510
	2	Green	510–580
	3	Red	630–690
	4	yellow	585–625
	5	Red edge	705–745
	6	Near-infrared 1	770–895
	7	Near-infrared 2	860–1040

In this study, 14 scenes were selected for model training and accuracy evaluation. The remaining data was reserved for visual evaluation. All the data was cropped into slices of size 256 × 256, and the 4400 slices of data obtained were divided into training samples. All data has been normalized and divided into two groups, with 70% and 30% used for training and validation, respectively. To promote transparency and reproducibility, the

data and code utilized in this research are made publicly available in the GitHub repository, which is linked at the end of the document.

3. Methods

3.1. Principle of CNNs for Land Cover Classification

In CNNs, convolutional layers are responsible for extracting local features from the image. These features are further processed through activation functions, such as ReLU, to introduce nonlinearity. Subsequently, pooling layers downsample the features to reduce computational load and increase invariance to image shifts. These layers typically alternate in the network until a fully connected layer (Fully Connected Layer) is reached, which maps the learned features to specific categories.

Assuming input data I has C_{in} channels, with each channel having dimensions $H_{in} \times W_{in}$, and the convolution kernel K has C_{out} output channels. Then, for each output channel at each pixel position (h, w), the convolution operation can be mathematically represented as:

$$O_{c_{out},h,w} = k_h \sum_{c_{in}=1}^{C_{in}} \sum_{k_h=1}^{K_h} \sum_{k_w=1}^{K_w} I_{c_{in},h+k_h-1,w+k_w-1} \cdot K_{c_{\text{out}},c_{in},k_h,k_w} \quad (1)$$

Here, $O_{c_{out},h,w}$ is the input image, $O_{c_{out},h,w}$ is the pixel value of the output feature map at (h, w, C_{out}) kernel, and $K_{c_{\text{out}},c_{in},k_h,k_w}$ is the weight of the convolutional kernel at the output channel C_{out}, input channel C_{in}, and kernel location k_h and k_w.

Land use classification tasks are typically multi-class classification problems, where each class corresponds to a type of land cover. As CNN learns from labeled examples, it leverages loss functions to adjust its weights and minimize the difference between its predictions and the actual labels. So, it is crucial for the spatial resolution of the data to be sufficiently high so that it can accurately classify a pixel based on its surrounding features, such as textures or shapes. Nevertheless, differentiating between shadows and water bodies, sparse vegetation areas and bare land, as well as buildings of varying materials and types, continues to pose a significant challenge. The augmentation of spectral information is indeed advantageous in resolving these issues. Speaking from the perspective of the convolutional computation process, leveraging multi-spectral data and increasing the number of input channels C_{in} enables the convolutional kernels to engage in a more diverse interaction and integration of features. This implies that the model can learn more complex feature representations, as each convolutional kernel weight $K_{c_{\text{out}},c_{in},k_h,k_w}$ can capture complex relationships between different channels. Although the number of parameters of the convolutional kernel weight $K_{c_{\text{out}},c_{in},k_h,k_w}$ does not increase, especially since its parameters are shared across all channels, using multispectral channels helps the model to learn richer features without significantly increasing the computational burden. Additionally, this operation is equivalent to expanding the model's receptive field in the spectral dimension.

This article utilizes two different loss functions: multi-class cross-entropy (CE) and weighted multi-class cross-entropy (WCE). The formulation of the multi-class cross-entropy loss function is:

$$J_{ce} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} log \hat{y}_{ij} \quad (2)$$

where n represents the number of samples, m denotes the number of the land cover classes, y_{ij} is the true label indicating if the i-th sample belongs to the j-th class and $\hat{y}_{ij}$ represents the predicted probability of the i-th sample belonging to the j-th class.

In the context of land cover classification, the weighted cross-entropy loss allows the model to learn more effectively from datasets where some land cover types may be less prevalent than others. By adjusting the loss function to penalize misclassifications of minority classes more heavily, the model can achieve better overall performance and produce more balanced classification results. The class w_j weights are calculated based on the inverse frequency of each class in the input data. This will ensure that the model is

optimized to handle the specific class imbalances present in the remote sensing data for land cover classification tasks.

The formulation of the WCE loss function is:

$$J_{wce} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} w_j y_{ij} log\hat{y}_{ij} \quad (3)$$

In this case, w_j represents the weight of the j-th class.

The workflow diagram of this paper is shown in Figure 2.

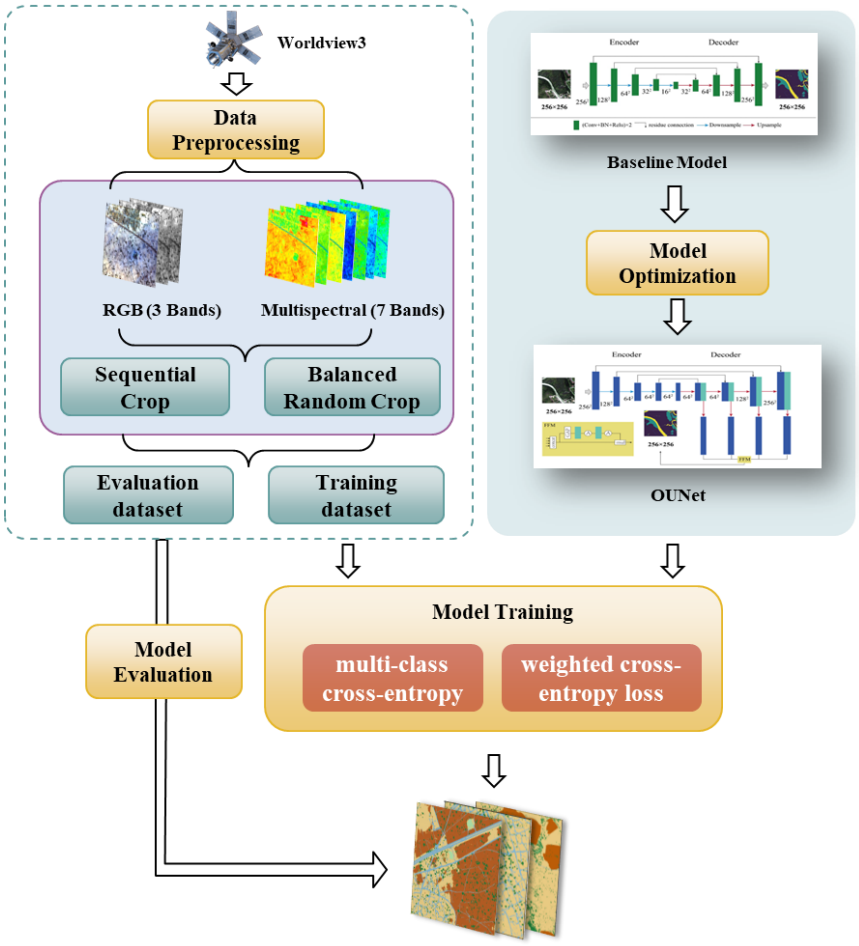

Figure 2. Workflow diagram of land cover classification.

3.2. Baseline Model

The UNet was proposed in 2015, and the network consists of an encoder and a decoder. The encoder includes several convolutional and pooling layers for feature extraction. In order to improve computational efficiency and expand the receptive field, four downsampling operations were performed on the encoder part. On the other hand, the decoder is composed of several convolutional and upsampling layers, which are used to restore the downsampled features and fuse them with same-scale features in the encoder to extract high-level semantic features, ultimately forming an end-to-end semantic segmentation network. The UNet's encoder is identical to the VGG16 [35] network, with the basic convolutional unit consisting of a 3×3 convolutional layer and an activation layer. The specific architecture of UNet is shown in Figure 3.

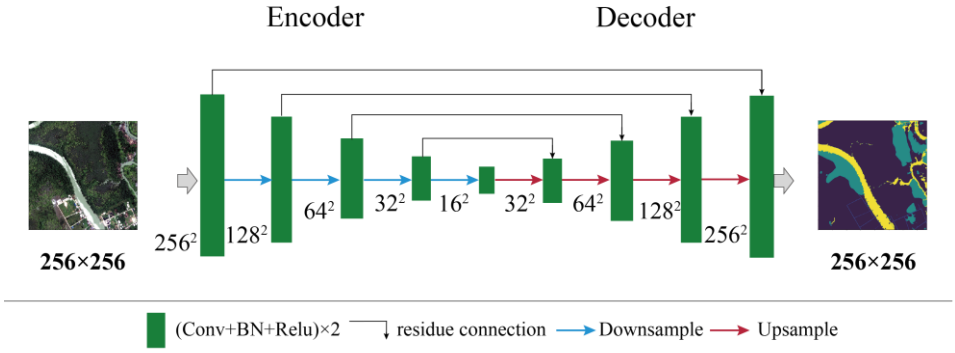

Figure 3. The architecture of UNet.

3.3. The Optimizing of Baseline Model

The UNet model has the advantages of fewer parameters and a simpler structure, which have made it widely used in computer vision, medicine, and other fields. However, there are still many aspects to be improved in the resolution of land cover classification based on high-resolution remote sensing. Therefore, we propose OUnet, which has made improvements in the following five aspects based on UNet: First, while retaining more shallow spatial information, the downsampling operation is reduced to improve the edge accuracy of remote sensing land classification results. Second, depthwise separable convolution is used to replace ordinary convolution to further reduce model parameters without significantly reducing model performance. Third, an attention mechanism is introduced in the decoder to better integrate features at different scales. Finally, a dropout layer is added to prevent overfitting of the model. While increasing the number of convolutions in the encoder effectively improves model performance, this article did not optimize OUnet in this aspect to facilitate better comparison with the baseline model. The baseline model is optimized using the above method to obtain OUnet, and its architecture is shown in Figure 4.

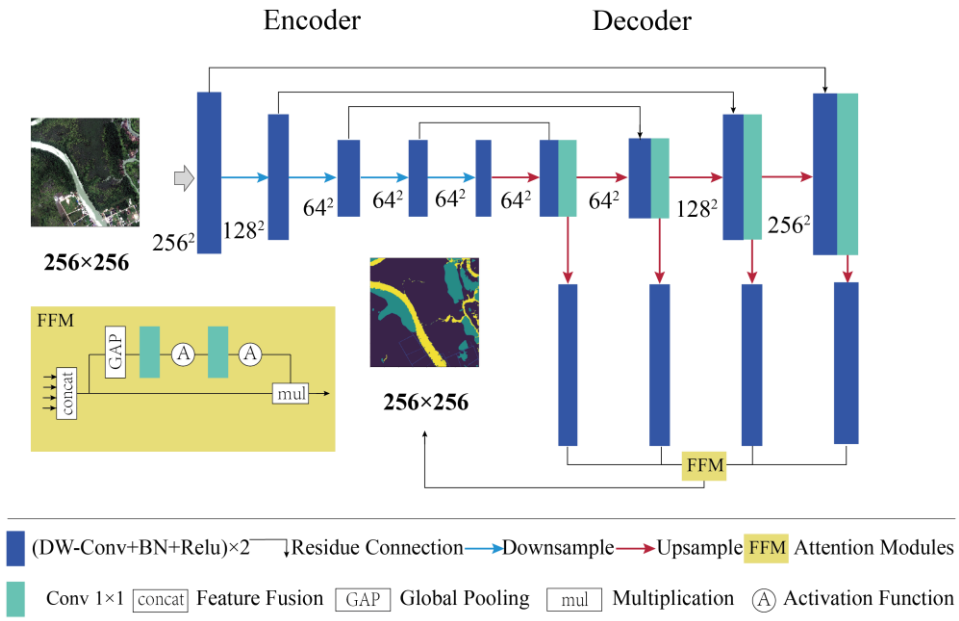

Figure 4. The architecture of OUNet.

3.4. Evaluation Metrics

In this paper, we utilize Mean Intersection over Union (*IoU*), mIoU, and the F1 score as comprehensive evaluation indexes for land cover classification. These metrics are pivotal in assessing the accuracy of the classified land cover types, reflecting the spatial congruence

between predicted and actual land cover maps. Precision and recall are also considered for a detailed evaluation of classification performance. The *IoU* for each land cover class is calculated as the ratio of true positives (*tp*)—pixels correctly classified as belonging to that class—to the total number of pixels that are either true positives, false positives (*fp*), or false negatives (*fn*):

$$IoU = \frac{tp}{fp + fn + tp} \tag{4}$$

The *mIoU*, a mean of IoUs across all classes (*k*), offers an aggregate measure of classification accuracy:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{tp}{fp + fn + tp} \tag{5}$$

The F1 score balances the trade-off between precision (the ratio of true positives to the total predicted positives) and recall (the ratio of true positives to the total actual positives), encapsulating the model's overall effectiveness:

$$\mathrm{F}_1 = 2 \,/\, \left(\frac{1}{\text{recall}} + \frac{1}{\text{precision}} \right) \tag{6}$$

where precision is given by:

$$precision = \frac{tp}{tp + fp} \tag{7}$$

And recall is given by:

$$recall = \frac{tp}{tp + fn} \tag{8}$$

Each metric provides a different perspective on the classification results: *IoU* and mIoU focus on spatial accuracy, while precision and recall provide insights into the types of errors made by the classification model, respectively.

4. Experimental Results and Discussion

4.1. Experimental Setup

Different data combinations, sampling methods, models, and loss choices may affect the classification results, and the goal of this study is to quantify such differences experimentally. We set up five experiments, as shown in Table 2. Two combinations of three-band true-color data and seven-band multispectral data were provided in the experiments. Although there have been studies that have basically clarified that the rich spectral information contained in multi-band data is more helpful for remote sensing land cover classification applications, it needs a more intuitive expression to determine whether the performance improvement is significant and whether the efficiency has significantly decreased. Two types of slice sample collection methods are provided in the experiment: one is sequential sampling, in which slices are cropped along the image length and width in fixed steps; the other is balanced sampling, in which slices are randomly generated within the image and the sampling balance is adjusted by limiting the proportion of each class in the labels. This approach is akin to performing data augmentation on the minority class, while random sampling ensures that the angles and scenes during sampling are dynamically varied, which further enhances the richness of the data. The loss function is compared using multi-class cross-entropy or weighted multi-class cross-entropy. It is worth noting that, to better illustrate the importance of the basic model optimization principle, we only compare the UNet and OUnet with the commonly used U-shaped framework. Although using a more complex structure (larger convolutional kernels, deeper networks) model will definitely improve the accuracy of experimental results, we did not do so in order to ensure fairness in comparison.

Table 2. Experimental settings.

Name	Model	Data Combination	Sampling Method	Loss Function
Baseline	UNet	Three-band	sequential	CE
Opt_1	OUNet	Three-band	sequential	CE
Opt_2	UNet	Seven-band	balanced	CE
Opt_3	UNet	Seven-band	balanced	WCE
Opt_4	OUNet	Seven-band	balanced	WCE

The experiment was conducted in a Windows 10 environment with a CPU of Gold 5218@2.3 GHz (×2), 256 GB of memory, and an NVIDIA Tesla A100 GPU. The deep learning framework used was TensorFlow (2.6.0). During the training process, the adaptive learning rate optimization algorithm was used as the optimizer, with an initial learning rate of 0.0001 for optimization. All models were trained for 80 epochs, and the best model among them was selected for comparison.

4.2. Experimental Results of the Baseline Experiments

In the baseline experiments, the UNet model is trained on a three-band dataset that was obtained using sequential sampling methods. Cross-entropy is used as the loss function during training. The validation results of the test dataset (Table 3) show that the classification accuracy is not high, with an mIoU of 0.52. There are significant differences in the classification accuracy of various land cover types. The accuracy for farmland and buildings is relatively high, with IoU both reaching above 0.75 and F1 scores both reaching above 0.85, while the accuracy for roads and water bodies is poor at 0.256 and 0.218, respectively. From the above data, it can be inferred that the strategy used in the baseline experiment has limited ability to extract multi-class land features.

Table 3. Performance of the baseline experiments.

Types	Accuracy	Recall	F1	IoU
Buildings	0.850	0.895	0.872	0.773
Roads	0.716	0.285	0.408	0.256
Trees	0.755	0.685	0.718	0.560
Farmland	0.863	0.904	0.883	0.791
Water bodies	0.466	0.291	0.358	0.218
Average	**0.730**	**0.612**	**0.648**	**0.520**

The Figure 5 illustrates the distribution of land cover types across the study area by representing the total area for each land cover category. The horizontal axis displays the area measurements, while the vertical axis lists the distinct land cover types. By analyzing the proportion of various land cover types in all samples, it is evident that there is a significant imbalance in the representation of different land cover types within the dataset. Notably, the categories of water bodies and roads are represented by the shortest bars, indicating that these classes have the smallest proportion of the total area sampled. This under-representation of certain land cover types in the dataset can lead to a scarcity of training examples for these classes. The limited number of training samples for water and road classes adversely affects the model's ability to learn and predict these land cover types effectively. Consequently, this scarcity is identified as a primary contributor to the observed lower accuracy rates in the classification of these specific land cover types. It is worth mentioning that, although the sample size of trees does not have an advantage compared to crops and buildings, the features of trees in remote sensing images are more distinguishable from the background compared to other land cover types, which results in a significantly higher classification accuracy than roads and water bodies. This also indirectly proves that using multi-band information as input data can provide the model

with more effective information favorable for classification compared to three-band data, thereby improving classification accuracy.

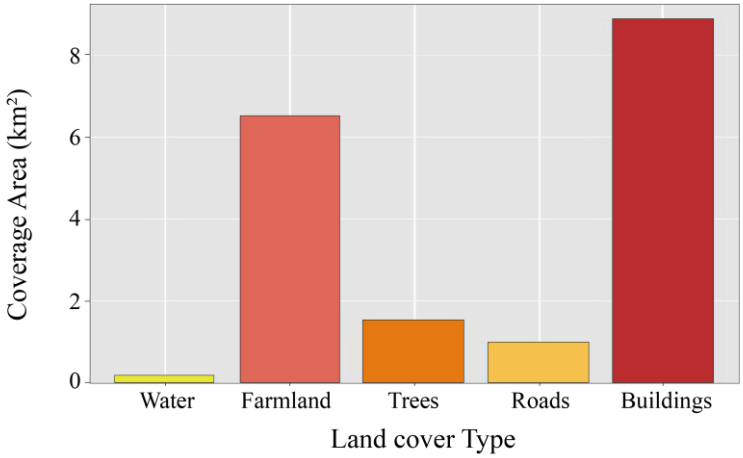

Figure 5. The distribution of various land cover types in the training data.

4.3. Optimization Experiments and Result Analysis

To address the issue of sample imbalance, two optimizations were strategically implemented, enhancing the statistical robustness of our approach. Initially, we employed random sampling to procure image slices for training, meticulously adjusting the proportion of each class in the labels to ensure a balanced representation across the dataset. Subsequently, we incorporated a weighted multi-class cross-entropy loss function, which assigns higher loss penalties to underrepresented categories, thereby compelling the model to pay closer attention to these samples during training. This function gives samples with lower proportions of each category higher loss penalties based on their actual proportions. This forces the model to learn more features from these types of samples.

To verify the effectiveness of the proposed optimization strategy, five sets of comparative experiments were performed for validation, with a particular focus on the precision of land cover mapping using high resolutions imagery (Table 2). Figure 6 illustrates the prediction results of different experiments. It can be observed that in the prediction results of the baseline experiment, water, trees, and roads, which constitute a relatively small proportion of the data, exhibit more pronounced misclassifications. However, this situation has been alleviated with the implementation of other optimization strategies. The experimental group Opt_1 achieved a notable enhancement in the identification accuracy across various land cover classes through optimization of the UNet model. However, misclassifications and omissions are still observed, particularly in the scenarios of water bodies and roads, which are less prevalent land cover categories. The Opt_2 experimental group further ameliorated this situation (Figure 6b,c). Despite no alterations to the model itself, the implementation of data augmentation strategies enriched the dataset, thereby enhancing the model's discriminative performance, although a noticeable gap with the ground truth remained. This suggests that the UNet model's capacity for land cover classification was not the primary cause of the suboptimal identification outcomes. The Opt_3 group incorporated a sample balancing sampling strategy and a weighted information entropy loss function, which further enhanced the recognition accuracy for small sample categories such as water and roads. This suggests that enriching the input data and balancing the sample distribution are highly effective in improving the multi-class land cover classification accuracy of remote sensing. The overall prediction results of the three experimental groups (Opt_2, Opt_3, and Opt_4) are relatively close to the ground truth, underscoring the effectiveness of our optimization strategies in enhancing the precision of land cover classification on high-resolution imagery. Among them, the Opt_4 group, which adopted

all optimization strategies, achieved the best performance, with the lowest incidence of misclassification and omission.

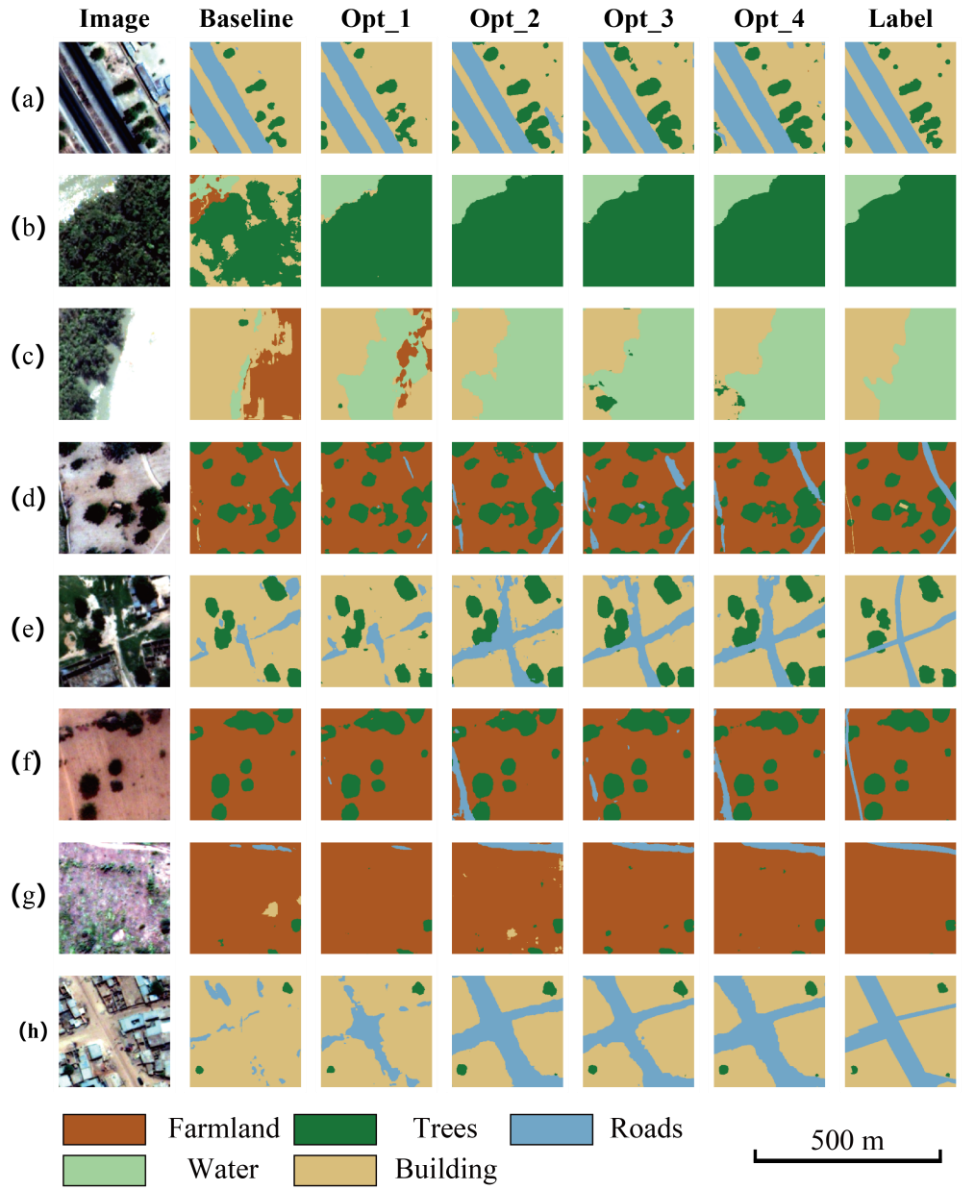

Figure 6. Classification results of different models. (**a**–**h**) represent image cases in different scenarios.

Table 4 presents the results of the quantitative analysis for all experiments, which align with the conclusions drawn from the preceding visual analysis. The four optimization experiments demonstrate the efficacy of the proposed method from various perspectives. The proposed improvement strategies prove effective for multi-class classification applications, resulting in enhancements of 29%, 17%, and 19% in recall, F1, and IoU, respectively, compared to the baseline experiment. Notably, the Opt_2 and Opt_3 groups exhibit higher accuracies than the Opt_1 groups, suggesting that augmenting the quantity and diversity of samples is more effective in mitigating accuracy degradation caused by sample imbalance in remote sensing multi-classification tasks than optimizing the model itself.

Table 4. Quantitative comparison results of all experimental groups.

Name	Precision	Recal	F1-Score	mIoU
Baseline	0.730	0.612	0.648	0.520
Opt_1	0.843	0.763	0.797	0.675
Opt_2	0.851	0.770	0.793	0.685
Opt_3	0.767	0.889	0.814	0.692
Opt_4	**0.777**	**0.909**	**0.825**	**0.709**

As shown in Table 4, the optimal experimental group 4, which integrated all proposed strategies, is markedly superior to that of the baseline group with a mean Intersection over Union is 0.709. This finding underscores the synergistic impact of enriching input data and balancing sample distribution on the precision of land cover classification. The recall of roads and water is improved by 66% and 70%, respectively, while the mIoU is improved by 30% and 58%, respectively (Table 5).

Table 5. Performance of the Opt_4 experiment.

Types	Accuracy	Recall	F1	IoU
Buildings	0.966	0.773	0.858	0.752
Roads	0.569	0.947	0.710	0.551
Trees	0.645	0.909	0.755	0.606
Farmland	0.910	0.918	0.914	0.842
Water bodies	0.797	0.996	0.885	0.794
Average	**0.777**	**0.909**	**0.825**	**0.709**

5. Discussion

5.1. Improve the Accuracy of Low Proportion Samples

In order to observe the impact of data imbalance on the classification accuracy in multi-class classification scenarios more carefully, we analyzed the recall rate, F1-score, and IoU index of the prediction results for the two least represented land cover types (roads and water bodies) in five sets of experiments. From Figure 7, it can be seen that all metrics in the baseline experimental group have the lowest values, while the Opt_3 experimental group has better results relative to Opt_2, suggesting that the weighted multi-class cross-entropy loss function is very effective for the data imbalance case for multiple landcover classification. By observing all the experimental data, it can be found that although the sample size of the water body category is small, its features are more distinguishable from the background compared to roads. Roads, as a typical linear shallow feature, spatial information is very important to improve recognition accuracy, which is taken into account in the design of OUNet. Therefore, the road accuracy improvement in the Opt_1 group is very obvious. When compared with Opt_3, it has limited room for improvement in sample conditions compared to data optimization.

Through the test comparison of the whole scene image (Figure 8), it can be observed that the improved method is closer to the true value, while the baseline method has obvious misclassification and omission phenomena in some of the feature recognition. Figure 9 complements Figure 8 by providing quantitative insights into classification metrics. It illustrates the significant improvements in classification metrics achieved by the optimized method over the baseline in multi-class land cover classification. The average mIoU has seen a notable enhancement for the optimized method with an improvement of approximately 44.03%. Individual category improvements are particularly striking in the road and water class, where the mIoU has marked an astounding increase, achieving over a 110% enhancement in both cases. These substantial gains underscore the optimized method's superior ability to accurately land cover features, with a marked reduction in misclassification and omission errors observed in the baseline approach.

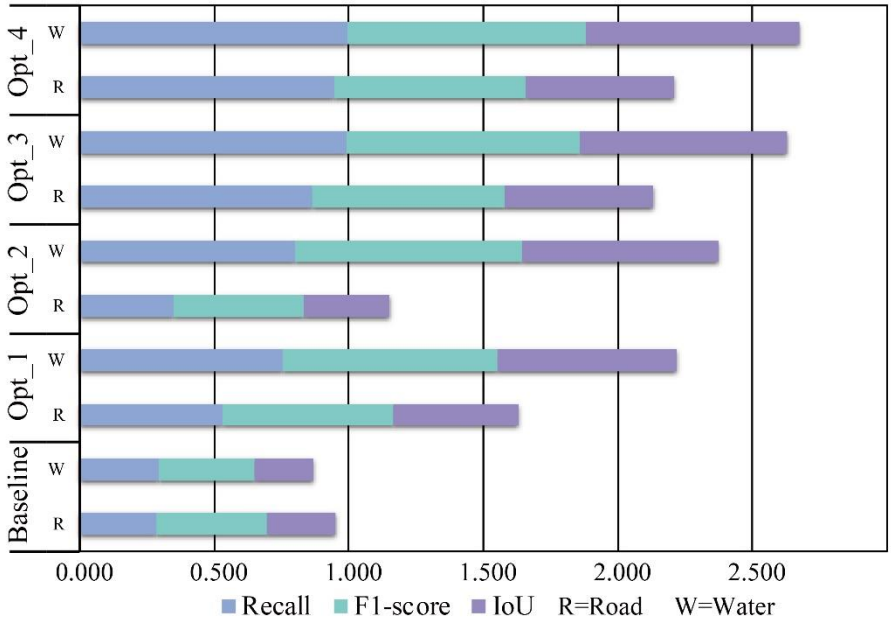

Figure 7. Comparison table of accuracy for roads and water in all experiments.

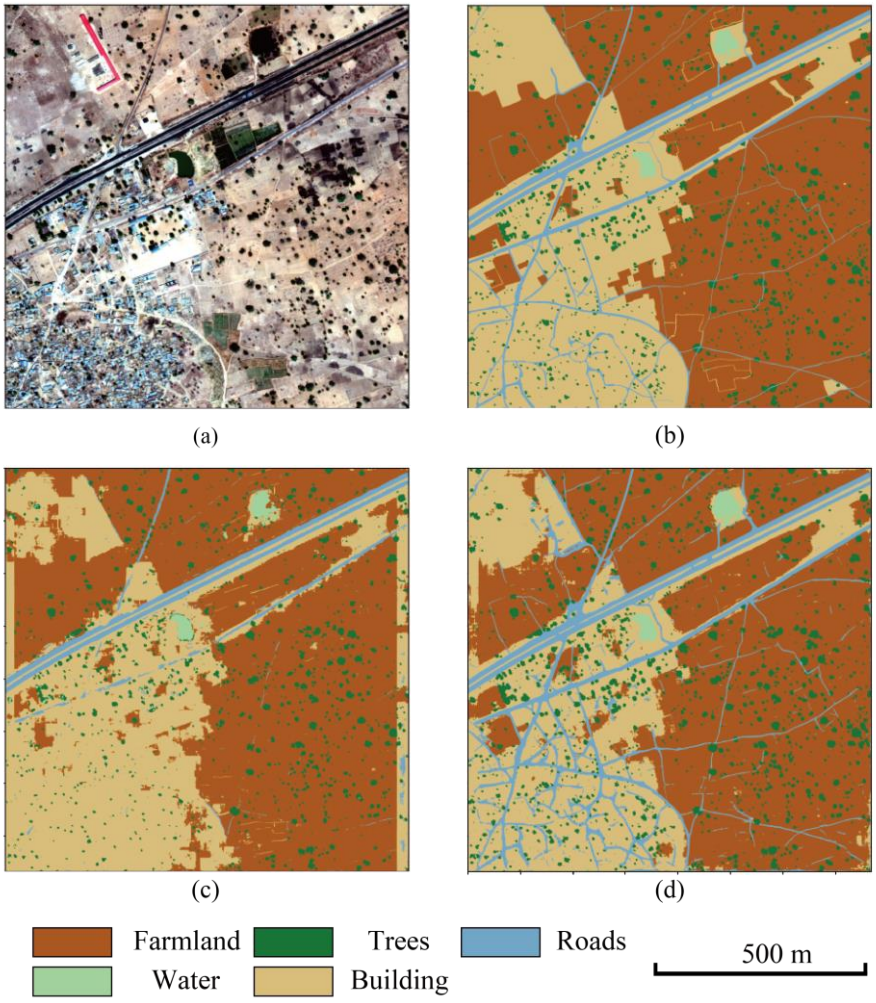

Figure 8. Performance of baseline and improved experimental results on test data. (**a**) RGB, (**b**) Reference tre value, (**c**) baseline solution prediction results, (**d**) Opt_4 solution prediction results.

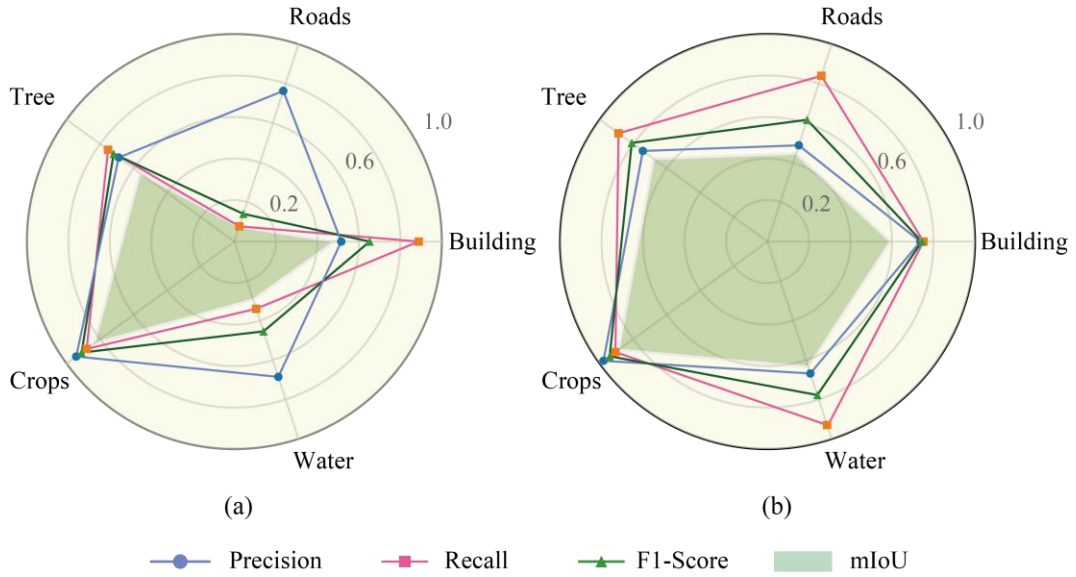

Figure 9. Performance evaluation radar chart. (**a**) Accuracy metrics for the baseline solution, (**b**) Accuracy metrics for the Opt_4 solution.

However, it should be noted that, while this paper has focused on illustrating the importance of addressing data imbalance for multi-class applications, the results presented in this research are amenable to further optimization. It can be further optimized by adopting a more sophisticated network, applying a diverse array of data augmentation strategies, and refining the inference process through advanced post-processing techniques and improved reasoning strategies. Nonetheless, the development of an accurate land use/land cover (LULC) product using deep learning technologies still necessitates a comprehensive consideration of ground truth conditions and the careful preservation of the topological features inherent in the landscape.

5.2. Superiority of Proposed Method

To substantiate the effectiveness of the optimization strategies proposed in this paper, we have employed two classical CNNs architectures, Dlinknet and DeepLabv3+, for further comparative analysis along dimensions such as prediction accuracy and model parameters scale. A first observation from Table 6 shows that in the experimental group using the baseline strategy, there is a direct correlation between model accuracy and parameter scale. On the other hand, in the Opt_3 experimental group, Dlinknet showed a 9% increase in performance without any increase in model parameters, due to optimizations in data and loss function, indicating the broad applicability of our proposed strategies to other CNNs as well.

Table 6. Comparation of CNNS models.

Types	Strategy	Precision	Recall	F1	mIoU	Params
UNet	Baseline	0.73	0.61	0.65	0.52	7.81
DlinkNet		0.80	0.67	0.71	0.58	25.32
Deeplab v3+		**0.81**	0.75	0.77	0.65	28.05
DlinkNet	Opt_3	0.75	0.87	0.79	0.67	25.32
UNet		0.77	0.89	0.81	0.69	7.81
OUNet	Opt_4	0.78	**0.91**	**0.83**	**0.71**	**6.57**

On another point, despite its superior structural design and larger parameter scale, DeepLabv3+ still underperforms compared to the optimized networks in Opt_3 in terms of performance gains. The higher accuracy but lower recall and mIoU values of DeepLabv3+

directly illustrate that while model optimization can improve the model's ability to identify positive samples, simply increasing model complexity does not address the issue of data imbalance, resulting in suboptimal recall and mIoU precision.

Finally, when examining the number of model parameters, the UNet using the Opt_4 strategy not only achieved the highest detection accuracy, but also had the lowest number of parameters. This is indicative of the cost-effectiveness of our proposed model optimization strategies, which can lead to a reduction in computational resources without compromising performance.

This paper introduces a novel multi-objective semantic segmentation framework tailored for high-resolution remote sensing imagery, aiming to enhance the accuracy and efficiency of multi-classification tasks. The proposed method leverages advanced data and model optimization techniques, which are theoretically extensible and could serve as a reference for the intelligent interpretation of a variety of remote sensing data types. Recognizing the prevalent challenge of small sample classification in practical applications, this study emphasizes the importance of harnessing the full potential of multispectral and hyperspectral data. To address the limitations inherent in traditional supervised learning paradigms, future research will focus on innovative unsupervised and semi-supervised learning strategies. These approaches will be pivotal in unlocking the value of small sample datasets, thereby facilitating more robust and generalized models for the remote sensing community.

6. Conclusions

In recent years, the application of deep learning models in the field of remote sensing has demonstrated significant potential for land cover mapping, particularly when leveraging high-resolution satellite imagery. In the realm of multi-class land cover classification, while there has been a notable emphasis on refining model structures to improve classification accuracy, it is imperative to also emphasize the pivotal role of data optimization and the strategic selection of loss functions. Both elements are integral to the enhancement of classification accuracy and should be considered as equally significant components of the overall optimization strategy. In this paper, the optimization method for multi-class land cover classification of high-resolution remote sensing imagery is investigated from different perspectives, such as sampling strategy, band combination, loss function, and model optimization. The findings of this research underscore the practical implications and the pivotal role of satellite imagery in achieving accurate land cover mapping. The main conclusions are as follows:

(1) This study reveals that high-resolution multispectral data is highly beneficial for distinguishing various types of land cover. Appropriate application of deep learning techniques enables CNNs to more effectively integrate spatial and spectral features, thereby mining the intrinsic value within the data and providing substantial assistance for land cover mapping.

(2) The experiments have demonstrated that the proposed optimization strategies, such as model optimization, data optimization through refined band combinations and sampling strategies, and loss function optimization, have effectively improved the richness of the data and mitigated the problem of sample imbalance. Consequently, the mIoU for land cover classification has been significantly increased from 0.52 to 0.71. Specifically, for low-occupancy land cover types such as roads and water, identification accuracy was remarkably improved by 30 and 58 percentage points, respectively. These results provide strong evidence for the effectiveness of the methods presented in this paper.

(3) In a comparative analysis with other CNNs methodologies, our proposed optimization strategies have proven to be equally efficacious when extended to architectures such as DlinkNet. The OUNet model, which was refined in accordance with our design principles, has not only attained the highest precision but has done so with a more streamlined parameter set. This not only underscores the effectiveness of our strategies

but also highlights their efficiency in terms of computational resources and model complexity.

(4) The models and data underpinning this study have been intentionally made accessible on GitHub, with a clear aim to furnish a practical and efficient tools for the land cover mapping application and remote sensing community.

With the rapid advancement of computer and artificial intelligence technologies, deep learning has been widely applied in land use and land cover (LULC) mapping, significantly enhancing automation and accuracy, especially for relatively homogeneous features. The recent rise of large remote sensing models and cloud computing further expands the possibilities in this field. However, we believe that a profound understanding of the application scenarios and objectives is a prerequisite for harnessing these new technologies. Fully exploiting the valuable information inherent in remote sensing data remains a crucial approach to problem-solving. Accurately positioning the role and boundaries of artificial intelligence within traditional practices can guide us towards more sustainable progress.

Author Contributions: Conceptualization, J.Y. and Y.L.; methodology, J.Y. and Y.L.; software, M.W.; validation, M.X. and L.X.; formal analysis, Y.P.; investigation, Y.L. and M.W.; resources, C.H.; data curation, M.X.; writing—original draft preparation, Y.L.; writing—review and editing, J.Y.; funding acquisition, J.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the National Key Research and Development Program of China (2021YFC3000400), the China Geological Survey Project (DD20243250, DD20242734), and the Major Special Project for High-Resolution Earth Observation System of Systems (04-H30G01-9001-20/22).

Data Availability Statement: The core algorithm of our study, the OUNet, has been implemented in TensorFlow and is now publicly available on GitHub at the following repository: https://github.com/JunchuanYu/OUNet_for_multi-class_land_cover_classification. Additionally, the training and testing datasets utilized in this paper will also be made accessible through the same GitHub repository.

Acknowledgments: We extend our gratitude to the HeyWhale Community (www.heywhale.com, (accessed on 15 March 2024)) for providing the computational platform. Additionally, we would like to thank Kaggle (www.kaggle.com, accessed on 1 March 2024) and the Defense Science and Technology Laboratory of the United Kingdom for the open-source dataset used in this paper.

Conflicts of Interest: The authors declare no conflict of interests.

References

1. Kemker, R.; Salvaggio, C.; Kanan, C. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 60–77. [CrossRef]
2. Lin, G.; Milan, A.; Shen, C.; Reid, I. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1925–1934.
3. Zheng, Z.; Zhong, Y.; Wang, J.; Ma, A. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. *Proc. IEEE/CVF Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2020**, *45*, 4096–4105.
4. Han, M.; Cong, R.; Li, X.; Fu, H.; Lei, J. Joint spatial-spectral hyperspectral image classification based on convolutional neural network. *Pattern Recognit. Lett.* **2020**, *130*, 38–45. [CrossRef]
5. Camps-Valls, G.; Bruzzone, L. Kernel-based methods for hyperspectral image classification. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 1351–1362. [CrossRef]
6. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
7. Yu, J.; Li, Y.; Zheng, X.; Zhong, Y.; He, P. An effective cloud detection method for Gaofen-5 images via deep learning. *Remote Sens.* **2020**, *12*, 2106. [CrossRef]
8. Zhan, Y.; Wang, J.; Shi, J.; Cheng, G.; Yao, L.; Sun, W. Distinguishing cloud and snow in satellite images via deep convolutional network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1785–1789. [CrossRef]
9. Hongqiang, Z.; Lingling, H.; Yongtian, W. Deep learning algorithm and its application in optics. *Infrared Laser Eng.* **2019**, *48*, 1226004. [CrossRef]
10. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]

12. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
13. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
14. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
15. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
16. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015, Proceedings, Part III*; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
17. Badrinarayanan, V.; Handa, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv* **2015**, arXiv:1505.07293.
18. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 10–14 September 2018; pp. 801–818.
19. Zhou, L.; Zhang, C.; Wu, M. D-LinkNet: LinkNet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 182–186.
20. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 10–14 September 2018; pp. 325–341.
21. Sun, K.; Zhao, Y.; Jiang, B.; Cheng, T.; Xiao, B.; Liu, D.; Mu, Y.; Wang, X.; Liu, W.; Wang, J. High-resolution representations for labeling pixels and regions. *arXiv* **2019**, arXiv:1904.04514.
22. Yuan, Y.; Chen, X.; Wang, J. Object-contextual representations for semantic segmentation. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part VI 16*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 173–190.
23. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12077–12090.
24. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning hierarchical features for scene labeling. *IEEE Trans. Pattern. Anal. Mach. Intell.* **2012**, *35*, 1915–1929. [CrossRef] [PubMed]
25. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
27. Chen, L.C.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A.L. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 3640–3649.
28. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3684–3692.
29. Wei, D.; Chen, J.; Luo, T.; Long, T.; Wang, H. Classification of crop pests based on multi-scale feature fusion. *Comput. Electron. Agric.* **2022**, *194*, 106736. [CrossRef]
30. Yu, J.; Zhang, L.; Li, Q.; Li, Y.; Huang, W.; Sun, Z.; Ma, Y.; He, P. 3D autoencoder algorithm for lithological mapping using ZY-1 02D hyperspectral imagery: A case study of Liuyuan region. *J. Appl. Remote Sens.* **2021**, *15*, 042610. [CrossRef]
31. Sahare, M.; Gupta, H. A review of multi-class classification for imbalanced data. *IJACR* **2012**, *2*, 160.
32. Rendon, E.; Alejo, R.; Castorena, C.; Isidro-Ortega, F.J.; Granda-Gutierrez, E.E. Data sampling methods to deal with the big data multi-class imbalance problem. *Appl. Sci.* **2020**, *10*, 1276. [CrossRef]
33. Dstl Satellite Imagery Feature Detection Competition. 2017. Available online: https://www.kaggle.com/c/dstl-satellite-imagery-feature-detection/data (accessed on 1 March 2024).
34. Iglovikov, V.; Mushinskiy, S.; Osin, V. Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. *arXiv* **2017**, arXiv:1706.06169.
35. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.